Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-691-AC3, 2017

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Value of seasonal streamflow forecasts in emergency response reservoir management" *by* Sean W. D. Turner et al.

Sean W. D. Turner et al.

stefano_galelli@sutd.edu.sg

Received and published: 20 April 2017

General Comments:

R: The paper concerns an interesting and emerging topic: quantifying the actual improvement of seasonal forecasts in water resources management and reservoir operation in particular. I think the experimental setting is valuable and provides interesting results, which are worth to be published, but I disagree about the way the results are presented, which may lead, in my opinion, to misleading interpretations. The paper distinguishes between "continually adjusted operation", when decisions are adjusted at frequent intervals, and "emergency response operation", when key decisions are taken infrequently. As an example of "continually adjusted operation", the paper considers a reservoir operated to track a constant target reservoir storage. As an example

of "emergency response operation", the papers consider a reservoir operated to meet a constant supply demand. The paper concludes that there is a clear relationship between forecast skill and value in the case of "continually adjusted operation" while this is not in the case for "emergency response operation". Another conclusion is that, in this second type of operation, it is of fundamental importance considering skillful forecasts at certain, well defined moments when critical decisions are taken. In my opinion, the results should not be commented in light of the type of decision process (in both cases decision are taken with the same frequency, i.e., every month), rather considering the type of reservoir operating objective and its dynamical response to forecasts. When the reservoir is operated for water supply, the release decision does not directly depend on the predicted inflow, because the reservoir storage partially buffers the variability of the inflows. As a consequence, poor inflow forecasts may not directly affect the operating performance (of course, this depends on reservoir capacity and storage-demand ratio). Prolonged poor forecasts (on long lead times, for example) may instead negatively affect the operating performance. When the reservoir is operated to track a certain target storage, the release decision must mimic the inflow much more closely (the reservoir storage does not provide any buffer in this case). For this reason, a skillful forecast implies high operating performance and viceversa, which turns into a clearer relationship between forecast skill and value. On top of this, we should consider that the storage tracking objective depends basically on the forecast of few lead times (if not on the first lead time only), while the supply objective is more dependent on forecasts on longer lead times (which allow for the hedging decisions mentioned in the paper). This means that, in general, the tracking storage operation benefits of more skillful forecasts than the supply objective, because both FoGSS forecasts and synthetic forecasts show a decreasing skill with lead time (Fig. 2 and 3). This may explain the fact that a certain forecast skill may be linked to a wider range of forecast values when considering the water supply objective (Fig. 5). I argue whether a different definition of forecast skill should be used in the two cases (for example, skill computed on different lead times, i.e., the ones relevant to the dynamic of each reservoir objective). I'll provide more

comments on these points in the following part.

Summarizing, I think that the paper deserve publication after a major revision. It might bring valuable insights on the topic of seasonal forecast value for reservoir operation, but the results should be discussed highlighting the effect of the different dynamics of the operating objectives (fast response to forecasts for storage tracking and slow response for water supply) on the forecast skill-value relationship, rather than focusing on the duality between "continually adjusted operation" and "emergency response operation", as it is in the current version of the manuscript.

A: *We greatly appreciate this detailed and thoughtful review. In particular we value the reviewer's the suggestion to distinguish the two reservoir types by operation rather than frequency of decision process. This shift in emphasis will allow for a clearer, more coherent interpretation of the results. We propose to make the distinction between the two reservoir objectives in these simple terms: one targets constant storage by varying the release, the other targets constant release by allowing the storage level to vary. This shift changes the role of storage from a target to a buffer, with consequent effects on forecast value that are brought out by our results. Rather than referring to the reservoirs as "continually-adjusted" and "emergency response", we'll use refer to "level objective" and "supply objective" (or similar). Discussion will place more emphasis on the role of a storage buffer in obviating the need for accurate forecasts during much of the operation.*

Specific comments:

R: Page 3, Lines 4-8. How is the sampling of the "error injected" parameter performed? Is there one sampling for each forecast? How does the error increase with lead time? The interpretation of the results would benefit by a more comprehensive description.

A: *Something we failed to mention in our paper was that the parameters of synthetic forecast model that define the way error increases with lead-time are defined using*

the actual FoGSS forecasts available for each catchment. This ensures that even though the forecast is synthetic, its decay with lead time is realistic. We sample the "injected error" once for each generated forecast. We will expand the description of the synthetic forecast model to clarify these details in our revision.

R: Page 3, Lines 25-26. "In months where forecasts are not informative, FoGSS is designed to return a climatological forecast." How is this performed? If this is the case, why does it happen that in Fig. 3 there are some negative skills (this is particularly evident for the Eppalock reservoir)?

A: *FoGSS applies statistical methods to correct errors in both the precipitation and hydrological components of the forecast, with the aim of ensuring forecasts are at least as skilful as climatology. The details are reasonably complex, but the principles are more straightforward. Essentially, precipitation forecasts can be considered as being corrected with a 'model output statistics' type approach (a term coined by Glahn and Lowry, 1972). (The actual method used, the Bayesian Joint Probability modelling approach, is described in detail by Wang et al. 2009, Wang and Robertson 2011, and as a post-processor of GCM forecasts, by Schepen and Wang 2014.) The principle here is that the relationship between the predictor, x, (e.g. forecast precip from a GCM for February at a lead-time of 3 months) and the predictand, y, (e.g. observation for that month) can be assumed to follow a regression, such that:*

$$y = dx + \mu \tag{1}$$

where d and μ are parameters. We optimise this relationship on the basis of historical forecasts, and where forecasts are consistently poor, d can approach zero. That is,

$$y \simeq \mu \tag{2}$$

This means that the forecast essentially returns a constant, μ (i.e., a climatology). In this way, precip forecasts should never be much worse than climatology (they can be slightly worse under cross-validation).

We pursue a similar approach for the hydrological component, with slightly different mechanics. In FoGSS, we completely separate uncertainty due to precipitation forecasts from uncertainty due to hydrological modelling. This means we do not apply a different regression to each lead-time, as we characterise hydrological model errors only from hydrological simulations that are forced by observations. That is, we apply a regression with parameters that vary by month, but these (12) sets of regression parameters are applied to all lead-times. This usually works quite well for forecasts (in concert with the precip corrections described above), but in instances where flow is often zero (as in Eppalock in late summer/early autumn), the hydrological model and the error model can tend to over-predict flows at longer lead-times. In the way that FoGSS is currently configured, the incidence of zero flows is always underestimated where observed flows are zero more than half the time, which results in small but persistent positive biases. While the actual discrepancies in flows in these cases are very small (e.g. perhaps the model predicts a mean flow of 0.6 mm, while climatological mean flow is 0.1 mm and, as already noted, observations are often zero), the skill (expressed as % of climatology flows, as in Figure 2) can appear to be very poor, because it is relative. We are currently working on methods to handle zero flows more proficiently, but this is a significant technical challenge. For the purposes of this paper, because negative skill occurs only when flows are very small, and forecast errors are extremely modest (often fractions of a mm), these issues have virtually no bearing on the utility of the forecasts to reservoir operation. We see this in our results where we demonstrate the strong value of forecasts at Eppalock.

We will add to the discussion of these issues in the paper as follows:

"In the Eppalock catchment, February and March usually experience very low (to zero) inflows. FoGSS forecasts in the Eppalock catchment are slightly positively biased

at longer lead times. Because flows are so low, even small small positive biases result in high relative errors in February and March. However, because inflows are so low during these months, these errors have very little influence on annual (or even seasonal) water balances." (Modified from the current text on Page 4, lines 3 - 5)

R: Page 4, Lines 24-26. In the standard operating policy, how is the release modulated when the demand cannot be fully met? I would include the description of this aspect also in Appendix 1.

A: *When there is insufficient water to meet the demand the release is simply constrained to the total volume of water left in storage plus available inflows. In other words, the operator releases as much as possible. We'll make sure this is clear in our revision.*

R: Page 4, Lines 26 and followings. How does the maximum release change in the different reservoir configurations? Is it linked to the reservoir capacity? Does the maximum release affect the ability of meeting the operating objectives (for example, should it be designed so to be able to always meet the demand)? The authors mention something in Appendix 2, but I would add a sentence in the main text as well.

A: *We simply set the maximum release to the demand multiplied by two for all systems, which means that larger capacity reservoirs have greater release capability. We think this assumption suffices for our synthetic study, although we concede that the release capability of actual reservoirs may vary widely depending on design. We'll follow the suggestion to add a sentence to the Appendix.*

R: Table 2. I think the readability of the table would benefit from a description of the meaning of "draft ratio", "drift" etc. in the caption. If I'm not wrong, "storage ratio" is not defined in the text. What does it represent? What are the actual reservoir features? It would be nice to have the figures for better understanding the reservoir settings in'

the different experiments. Is the "critical period" computed assuming no inflow to the reservoir or climatology or what else? The numbers seem in some cases very high and I don't fully understand how to interpret these large numbers (see also the following comment).

A: *We'll add some definitions to the caption, as suggested. The storage ratio (years) is simply mean annual inflow divided by the storage volume. It represents the number of years it would take to fill the reservoir if there was no release—so it provides some indication of reservoir's potential recharge rate. The critical period is computed using the historical inflow record. Since the reservoirs are designed to 0.95 reliability under the historical record, they are guaranteed to fail when operated with standard operating policy. So all simulations contain a period from full to empty that is used to determine the critical period. The critical period can be very long because often the drawdown rate is moderated by inflows. In particularly large systems with low recharge rates, it's often the case that the inflows will partially recover the reservoir before drawdown commences again. This can lead to very a long critical period.*

R: Figure 4. "Emergency response" case: why do the Burrinjuck and Eppalock reservoir empty, if their critical periods are 84 and 102 months respectively? "Continually adjusted" case: why does it happen that the reservoir storage is above the target but the releases are equal to 0? (see for example, Burrinjick and Eppalock reservoirs at the beginning of the time series) This behavior seems to be sub-optimal, because a release greater than zero would contribute in reducing the objective cost.

A: *All reservoirs are designed with a reliability of less than 1, so emptying is guaranteed (no emptying would mean no requirement to cut back demand, and so the reliability would be 100%). In figure 4 we observe that even with an extremely long critical period (102 months in the case of Eppalock) the reservoir does in fact empty toward the end of the drought period (taking precisely 102 months).*

The question regarding releases being zero despite high storage is very interesting. It most cases we observe that when the storage is above target the release will be large (because the objective is to bring the storage back down to target). However, there are some instances in which the model appears indifferent to the large storage level (such as the beginning of some of the time series, as identified by the reviewer). The reason is that in these instances the impact of zero release or full release is the exact same. The inflow in these instances is so large that even when maximum release is applied, the reservoir will continue to spill. So deviation from the objective function is the same irrespective of the release decision. Our model takes the lowest release as default in these instances, but because the reservoir would spill anyway the memory of the decision is wiped out and has no bearing on the end cost.

R: Page 6, Lines 13-14. "We test two operating objectives: one that rewards a judicious response to an emergency (emergency response objective) response and one that rewards judicious continual adjustments (continual adjustment objective)". As already mentioned, I don't agree with this terminology because decisions are taken at a constant pace when considering both the objectives. The "emergency response objective" could easily become a "continual adjustment objective" if, for example, the demand was (relatively) large and the reservoir (relatively) small, because the release decision could change much more frequently. Viceversa, the "continual adjustment objective" in Fig. 4 - Eppalock reservoir behave as an "emergency response objective", because the release decision is most of the time equal to the maximum release and does not change frequently, just because the storage is for a long time higher than the target. I would suggest changing this terminology because it is misleading in my opinion... Page 7, Lines 1-2. "For the continually adjusted operating setting we find that the release must be adjusted constantly through the operating horizon to keep storage close to the target level of 75%". As mentioned before, this objective is directly influenced by the variability of the inflow and reservoir release should change frequently to keep a constant storage. An example can be found in Fig.4 - Eppalock reservoir, where the

storage is at the target storage, but the release keeps changing because of the inflow to the reservoir. I would include this explanation in the text.

A: *We agree with this critique of the terminology and will reconstruct the paper consistent with reviewer's recommendation to compare the reservoir settings by focusing on the objectives and consequent role of storage rather than frequency of decision adjustment.*

R: Page 7, Lines 24-29. I would appreciate a discussion on the reason why the two objectives behave differently. The interpretation I propose in the "General comment" may represent a possible interpretation.

A: *See previous response.*

R: Page 7, Lines 29-32. I am not sure that the two objectives can be compared just on the basis of the value of the "injected forecast error". In fact, as commented in the "General comment", the "continually adjusted objective" is mainly driven by the forecast at one or few lead times, while the "emergency response objective" is driven by forecasts on longer lead times. On these lead times, the forecast skill is poorer by construction (if I understood how the synthetic forecasts were produced). What is the authors' opinion on this? Would it make sense to compute the forecast skill on different lead times, e.g. short lead times for the first objective and long lead times for the other objective? If, for example, it would turn out that the forecast skill in Fig. 5 a-d for injected error equal to 0.2 is comparable to the forecast skill of Fig.5 e-h for injected error equal to 0.6, spread of the forecast value would be similar in the two cases.

A: *Whilst we accept that reliance on forecasts at longer lead times will impair the cost reduction achieved, we expect the influence of lead time on the spread of performance to be modest. The reason is that the spread of forecast skill for a given injected error is similarly tight across all lead times (fig. 2). Nonetheless, we think this point is*

worth raising in the description, even if only to alert the reader to the point that longer lead-time forecasts are likely to be more important in the supply targeted reservoirs. The use of the longer lead time forecasts for the supply targeted reservoirs (upper panels of Figure 5) explains the better cost improvements with the perfect forecast and the much sharper drop off in performance when error is injected.

R: Page 7, Lines 33-34. Why do the Burrinjuck and Serpentine reservoirs behave differently from the other two reservoirs? What are the features that they have in common which may justify the observed behavior? I would be nice if this result could be generalized.

A: *This is actually a misstatement—thanks for picking this up. Eppalock is also sensitive to the increase in error, so it is Upper Yarra that is the outlier here. What we should have stated was: "Upper Yarra tolerates greater increases in error injected before forecast-based operations begin to be outperformed by control rules". We suspect the main factor at play here is the very short critical period for Upper Yarra relative to the other reservoirs (when draft ratio is 0.5). This means the storage buffer is less influential and adjustments to the release decision are required more frequently (i.e., it behaves a little more like the "continually adjusted" reservoir), as indicated in Figure 4. We'll expand this section to include some reasoning for the behavior observed.*

R: Page 7, Lines 38 and followings. "These results show that the measure of forecast error, quality, skill or goodness-of-fit—if based on the entire forecast period—cannot predict accurately whether that forecast will be valuable in an emergency-type operational setting". What do you mean when you write: "if based on the entire forecast period"?

A: *Here we're referring to measures of forecast quality that are based on an assess-*

ment of the hindcast as compared to observed conditions, over, say, 30 years. When we say "entire forecast period" we mean the skill measured across the hindcast (as opposed to selecting points within the hindcast, e.g., during drought, when the forecast performance may be more important to the performance in operations). We'll clarify this in the re-write, as it's a critically important point.

R: Page 8, Lines 2-4. "This may be because the emergency response objective is constructed to be sensitive to a few serious shortfalls in meeting demand, while the continual adjusted objective rewards consistent performance over all the months assessed". Both the objectives have a squared term, which should penalize to the same extent big deviations from the target (It could be different if there was an absolute value in Eq. 5, instead of the squared term). I think that the explanation may rely in the buffer effect of the reservoir, as explained in the "General comment".

A: *We agree that the explanation relies on the buffer effect of the reservoir to the extent that when reservoir is full the forecast is made redundant (the decision to release will be made irrespective of forecast quality). The result of this phenomenon is that the available improvements from forecasts in large supply targeted reservoirs depend solely on the quality during isolated periods when the reservoirs are drawn down. So (referring to the previous comment) measuring forecast performance across the whole hindcast may not indicate the value that will be reaped by that forecast in operation.*

R: Page 8, Lines 15-19. This sentence is not fully clear to me.

A: *We will the replace this section with: "Operations are then simulated using both the control rules and the deterministic model predictive control model using the median value from the full FoGSS forecast ensemble. (i.e., a deterministic forecast is constructed by taking the median of the ensemble at each lead time.) While this ignores the spread of the ensemble, the chosen method provides a clear indication*

of the contribution of the forecast to the performance of the operation. In contrast, methods that use the spread of the ensemble in the decision process are complex, often requiring arbitrary decisions by the user. This makes experimentation laborious and results hard to diagnose. . ." [example of multi-stage stochastic optimization with recourse].

R: Page 8, Line 25. "our own prior experiments with this approach". Please provide a reference, if any.

A: *We don't have a reference here. We'll notify the reader that the statement relates to unpublished work or simply remove the statement.*

R: Page 8, Line 35. I would explicitly state that the second experiment focuses mainly on the "emergency response objective".

A: *Agreed.*

R: Figure 6. I would include a marker on the left hand side figures. I was surprised to see that there are more critical periods in case of low draft ratios (see fig. a, c, e). I would expect that higher draft ratios drive more critical situations. Could the authors comment on this?

A: *It's important to remember that higher draft ratio reservoirs also coincide with larger storages (since all reservoir storages here are designed for reliability of 0.95). A general pattern that emerges in such circumstances is that reservoirs with larger demand (and storage) recover less readily, leading to a concentration of the 5% of time with failure on a single drought period. Smaller reservoirs can fail but then recover quickly, so the 5% of failure periods tend to occur multiple times over the simulation period. We'll add a sentence in the paper to notify the reader that this is the expected behavior.*

R: Page 9, Line 10-12. "since the reservoir is drawn down at the end of the simulation, meaning the implications of late, sacrificial decisions are unavailable to quantify overall performance". Does this mean that the penalty on the final storage is not considered in the optimization? If so, why is it not considered?

A: *The penalty of the final period is considered in the optimization undertaken for each forecast available. The issue here is that the impacts of late decisions for the full simulation period can't actually be compared fairly using a final period storage penalty. Clearly, ending with a higher storage would imply less cost using the storage penalty. But if the reservoir would recover anyway (if, for example, the inflows immediately after the simulation period were high) then the smarter action would have been to allow the storage to deplete more and meet the demand in full. In such an instance, the final storage penalty would misrepresent the value of the decisions taken leading up to the end of the simulation. In other words, we can't fairly evaluate the decisions taken within a given drawdown period unless the drought is allowed to play out in full (particularly when we're using lead times of up to one year).*

R: Page 9, Line 28. "the hedge comes too late at the end of 2006". This is currently not visible in Fig. 8. I wonder if including the trajectories obtained with the perfect forecast might clarify what is "too late".

A: *We agree that this is perhaps unclear from the figure presented, but we think that including the perfect forecast may make this plot overcomplex. Instead we'll annotate the figure to highlight the points raised.*

R: Page 9, Line 31-32. I don't understand the comment. The reservoir does not fully recover in both the trajectories. This might be because the inflow is too small in comparison of the reservoir storage.

A: *We'll aim to clarify this in our re-write. It is true that neither trajectory recovers. What this means is that the impact of earlier release decisions matters for both operating modes. The control rules hedged (correctly), whilst the forecast-based operation didn't—and because neither reservoir recovers, these decisions become important further into the future, resulting in weaker forecast-based operations relative to the control rules.*

R: Page 9, Lines 41-42. "forecast skill must be consistently available". "Consistently" means "on long lead times"?

A: *See earlier comment for Page 7, Lines 38. This is something we will define more clearly in our revision. By 'consistently', we do not mean "on long lead times". We mean that the skill of the forecasts, measured across, say, 30 years, should be consistently good if improved performance in operation is to be achieved. Forecast skill is usually calculated by averaging error scores over long periods, which can emphasise single forecasts that were either very good or very poor (relative to climatology). For example, a large majority of forecasts may perform slightly worse than climatology, but one or two forecasts perform very well, resulting in a positively skillful forecast (on average). However, in this example, the forecasts are unlikely to be useful to reservoir managers, because skill is not consistently available. This is particularly the case in supply reservoirs with large storage buffering capacity because the forecasts are seldom relied upon.*

R: Page 10, Lines 14 and following. I would revise the conclusion (as well as the abstract and title) to soften (or remove) the distinction between the "continually adjusted operation" and "emergency response operation", as already commented several times in this review.

A: *We'll amend the terminology used to contrast the reservoir operating settings in*

line with the reviewer's earlier suggestion (see response to the first reviewer comment).

R: Page 10, Lines 19-20. It seems that the results of the second experiment are driven mostly by the Millennium Drought. This exceptional sequence of several dry years contributes in enhancing the value of forecasts. Would the results be much different if his extremely exceptional event was not considered? It would be interesting to have a look at the figures computed without the last years, to have an idea of the forecast value in case of less extreme droughts.

A: *In forecast verification, we wish to understand how forecasts are likely to perform in future. This is best achieved by looking at forecast performance for the longest periods possible, making it undesirable to shorten the verification period. In addition, as future observations cannot be known with certainty, it only makes sense to condition verification on forecasts of dry periods (not observations). This is obviously not possible for multi-year droughts, because our forecasts only predict out to 12-months.*

Technical corrections

R: Page 6, Equation 4-5. Should "T" be "H", given the notation in Eq. 3? Define all the variables, not only "D".

A: *H is the 12 month forecast horizon, whereas T is the entire simulation period. We'll clarify this point and define all variables, as suggested.*

R: Page 8, Line 12. "24 reservoirs" should be "32 reservoirs".

A: *Will update.*

R: Page 9, Line 20. Fig. 8 seems to be cited before Fig. 7 in the text.

A: *Will correct.*

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-691, 2017.